Estimating Heterogeneous Effects in Binary Response Models

PRELIMINARY AND INCOMPLETE. COMMENTS ARE WELCOME.

Anastasia Semykina
Department of Economics
Florida State University
Tallahassee, FL 32306-2180, USA
E-mail: asemykina@fsu.edu
Phone: +1 850-644-4557
Fax: +1 850-644-4535

October 20, 2017

Abstract

The paper considers estimating heterogeneous effects on binary outcomes in different population subgroups. Applications include the evaluation of heterogeneous treatment effects, as well as estimation of causal effects of other policy-relevant variables. In the existing literature, it is common to divide the sample into group-specific subsamples and perform estimation separately for each group. In this paper, we argue that this estimation approach generally results in inconsistent estimators and present estimation methods that produce consistent estimators of causal effects in heterogeneous populations. The theoretical argument is illustrated with an empirical example.

JEL Classifications: C31, C33, C35
Keywords: binary response, heterogeneous effects, nonrandom sorting

1 Introduction

Researchers are frequently interested in estimating heterogeneous effects on binary outcomes in different population subgroups. Examples include studying dropout rates among high school students by gender and race, examining labor market participation decisions among married and single women, and investigating self-employment outcomes by age and education level. In the empirical literature, it is common to estimate such group-specific parameters by dividing the sample into the corresponding subsamples and performing the estimation separately for each group. While this approach is intuitively appealing, we argue that it generally results in inconsistent estimators when sorting into groups is not random. Moreover, as shown by Vella (1988) in the context of linear models, the consistent estimators of heterogeneous parameters can only be obtained if the full information set is utilized, i.e. when each group is considered as a part of the entire population. The present paper discusses the estimation procedures that address the mentioned problems and produce consistent estimators of heterogeneous parameters in binary response models.

The models considered in this paper are related to the literature on the linear switching regression models (Goldfeld and Quandt, 1973; Lee 1978; Maddala and Nelson, 1975; Maddala 1983). The switching regression models specify two equations, where the applicability of either equation depends on the endogenous switching from one regime to the other. Another relevant strand of the literature includes studies on program evaluation and estimation of treatment effects. Analogous to the switching regression models, program evaluation literature is focused on addressing the endogenous self-selection into treatment. One parameter of interest in these studies is the effect of the treatment on the treated, which can be formulated within a switching regression or self-selection framework (Bjorklund and Moffitt, 1987; Heckman et al., 2006). Furthermore, several studies have proposed methods for estimating heterogeneous treatment effects using the instrumental variables methodology (Heckman et al., 2006; Basu, 2014, and others).

The problem of nonrandom selection is also discussed in the studies of sample selection, including the seminal paper by Heckman (1979). In those models parameters are assumed to be the same for all units in the population, and the selection problem arises because the value of the dependent variable is not observed for some part of the population. In the previous literature, the methods for addressing the sample selection in linear and binary response models were proposed. The estimators were developed to address the selection problem in both cross section and panel data models (Heckman, 1979; Kyriazidou, 1997; Newey, 2009; Semykina and Wooldridge, 2017; Wooldridge, 1995, among others).

Considering heterogeneity in binary response models, several studies discuss the switching probit model for cross section and panel data (Carrasco, 2001; Manski et al., 1992). Similar to linear models, the endogenous switching is between two regimes, and parameters are regime-specific. However, to the best of our knowledge, estimating general heterogeneous effects models with an arbitrary number of groups (or regimes) has not been considered so far. To goal of this paper is to fill the gap in the literature by presenting the methods for estimating heterogeneous effects in binary response models with two or more groups. In the presented discussion, we distinguish between the models where the groups are ordered and models with unordered multiple groups.

The rest of the paper is structured as follows. Section 2 presents binary response models with heterogeneous effects. Estimation of the model parameters is discussed in Section 3. Section 5 contains an empirical application, and Section 6 concludes.

2 Heterogeneity in binary response models

Consider a population that consists of J groups (or subpopulations) and write a binary response model with heterogeneous effects as

$$\begin{aligned} y_j^* &= x\beta_j + u_j, \\ y_j &= 1[x\beta_j + u_j > 0], \qquad j = 1, \ldots, J, \end{aligned} \tag{1}$$

where y_j^* is a latent variable, and y_j is the observed outcome in group j, $1[\cdot]$ is an indicator function equal to one if the expression in brackets is true, x is a $1 \times K$ vector of explanatory variables that is independent of error u_j, and β_j is a vector of parameters that is different for each group j. For each j, β_j is considered to be fixed, so that it is independent of x and u_j.

Let d be a discrete random variable identifying groups, $d = \{1, 2, \ldots, J\}$. After defining dichotomous indicators for each group as $s_j = 1[d = j]$, $j = 1, \ldots, J$, the outcome in the entire population can be written as

$$y = \sum_{j=1}^{J} s_j y_j. \tag{2}$$

In this paper, we focus on the case, where the primary interest is in estimating the partial effects of covariates in each group with the ultimate goal of performing comparisons across groups. For example, for a continuous explanatory variable x_k, one would like to estimate

$$PE_{j,k} = \frac{\partial \mathrm{P}(y = 1|d = j, x)}{\partial x_k} = \frac{\partial \mathrm{P}(y_j = 1|d = j, x)}{\partial x_k}, \qquad j = 1, \ldots, J, \tag{3}$$

where $PE_{j,k}$ is the partial effect of x_k conditional on being in group j. If group assignment is random, then consistent estimators of β_j and PE_j can be obtained by estimating

equation (1) separately for each group and computing the partial effects within that group. However, because of self-selection or other factors, sorting into groups may be nonrandom, which generally leads to inconsistency. In this paper, we allow for a possibility that $\mathrm{P}(d = j|u) \neq \mathrm{P}(d = j)$ and discuss how it can be addressed in the estimation of β_j and partial effects. We start by considering a simple case with only two groups and then discuss more general models with $J > 2$, where groups may be either ordered or unordered.

2.1 Model for two groups

Let y_j be determined as in equation (1), with $J = 2$. The applications of such models could include, for example, examining labor force participation among married and non-married women, as well as estimating the determinants of dropout incidents among economically disadvantaged and other students. Assume that sorting into groups is determined by the value of latent variable d^*,

$$\begin{aligned} d^* &= z\delta + v, \\ d &= 1 \text{ if } d^* \leq 0, \\ d &= 2 \text{ if } d^* > 0. \end{aligned} \tag{4}$$

where z is a $1 \times L$ vector of exogenous variables, and v is the error term. Setting a cutoff point at zero is at no cost, as long as z contains an intercept. Let y_j, $j = 1, 2$, and d be defined as in equations (1) and (4), respectively. Also, assume that the following holds:

ASSUMPTION 2.1 (i) (u_j, v) are independent of (x, z), $j = 1, 2$, (ii) $z = (x, z_1)$, where z_1 is not empty, (iii) (u_j, v) have a bivariate normal distribution with $\mathrm{Var}(u_j) = \mathrm{Var}(v) = 1$ and $\mathrm{Corr}(u_j, v) = \rho_j$, $j = 1, 2$, (iv) $0 < \mathrm{P}(d = j) < 1$, $j = 1, 2$.

Assumption 2.1(i) is a standard exogeneity condition that implies that inconsistencies

in group-by-group estimation may result only due to nonrandom sorting (or self-selection), but not because of endogenous explanatory variables. The second assumption, 2.1(ii), requires that z contain x and at least one more variable, which is needed for identification. Without 2.1(ii), β_j can only be estimated based on the nonlinearity of the likelihood function, which is undesirable. The normality assumption, 2.1(iii), is rather standard in the literature and permits obtaining formulae for the conditional probabilities and partial effects. Finally, the last part of assumption 2.1 ensures that there are cross section units in each group.

Notice that under the specified assumptions, the two-group model is a switching probit model, which is analogous to a linear switching regression model discussed in the literature (Carrasco, 2001; Lee 1978; Maddala and Nelson, 1975; Maddala 1983; Manski et al., 1992). It is evident that when $\rho_j = 0$, sorting is completely random and can be safely ignored because

$$\mathrm{P}(y = 1|d = j, z) = \frac{\mathrm{P}(y_j = 1, d = j|z)}{\mathrm{P}(d = j|z)} = \mathrm{P}(y_j = 1|x), \qquad j = 1, 2. \tag{5}$$

In a general case, however, one has to account for a possibility of nonrandom sorting or self-selection. In the linear switching regression model, it is usually addressed by constructing a correction term that captures the conditional expected value of u_j given $(v, d = j)$. In binary response models, the nonlinearity of the conditional mean makes such correction impossible. Instead, one has to consider the conditional distribution of u_j given $(v, d = j)$. Under Assumption 2.1, using the properties of normal distributions we can write

$$\begin{aligned} u_j &= \rho_j v + e_j, \\ e_j|z, v &\sim Normal(1, 1 - \rho_j^2), \end{aligned} \tag{6}$$

so that the dependent variable for group j can be written as

$$y_j = 1[x\beta_j + \rho_j v + e_j > 0], \qquad j = 1, 2. \tag{7}$$

Then, the conditional probability for the first group can be written as

$$\begin{aligned} \mathrm{P}(y = 1|d = 1, z) &= \frac{\mathrm{P}(-e < x\beta_1 + \rho_1 v, v \leq -z\delta|z)}{\mathrm{P}(v \leq -z\delta|z)} \\ &= \frac{\int_{-\infty}^{-z\delta} \Phi\left(\frac{x\beta_1+\rho_1 v}{\sqrt{1-\rho_1^2}}\right)\phi(v)dv}{1 - \Phi(z\delta)}, \end{aligned} \tag{8}$$

and the corresponding conditional probability for the second group is

$$\mathrm{P}(y = 1|d = 2, z) = \frac{\int_{-\infty}^{z\delta} \Phi\left(\frac{x\beta_2+\rho_2 v}{\sqrt{1-\rho_2^2}}\right)\phi(v)dv}{\Phi(z\delta)}, \tag{9}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution functions, respectively.

The partial effects of continuous explanatory variables can be obtained by differentiating the probability function in (8) or (9) with respect to x_k. Specifically, for $j = 2$,

$$\begin{aligned} PE_{2,k} = \frac{\partial \mathrm{P}(y = 1|d = 2, z)}{\partial x_k} &= \delta_k \cdot \frac{\phi(z\delta)}{\Phi(z\delta)} \cdot \Phi\left(\frac{x\beta_2 + \rho_2 z\delta}{\sqrt{1-\rho_2^2}}\right) \\ + \frac{1}{\Phi(z\delta)} \cdot \frac{\beta_{2k}}{\sqrt{1-\rho_2^2}} &\cdot \int_{-\infty}^{z\delta} \phi\left(\frac{x\beta_2 + \rho_2 v}{\sqrt{1-\rho_2^2}}\right)\phi(v)dv - \delta_k \cdot \frac{\phi(z\delta)}{\Phi(z\delta)} \cdot \mathrm{P}(y = 1|d = 2, z), \end{aligned} \tag{10}$$

where $\mathrm{P}(y = 1|d = 2, z)$ is as defined in equation (9). For $j = 1$, the partial effects are obtained by replacing δ (δ_k) with $-\delta$ ($-\delta_k$) and changing β and ρ subscripts to one. Notice that when the group assignment is random, the partial effects on the conditional probabilities are the same as the unconditional partial effects. However, they are different when $\rho_j \neq 0$.

The partial effects of discrete variables (such as binary indicators) are obtained by considering changes in conditional probabilities. For example, for a discrete variable h,

$$PE_{j,h} = \mathrm{P}(y|d=j,z^1) - \mathrm{P}(y|d=j,z^0), j=1,2, \tag{11}$$

for $z^1=(x^1,z_1)$, $x^1=(x_1,\ldots,x_{h-1},x_h^1,x_{h+1},\ldots,x_k)$ and $z^0=(x^0,z_1)$, $x^0=(x_1,\ldots,x_{h-1}$, $x_h^0,x_{h+1},\ldots,x_k,)$.

2.2 Model for multiple ordered groups

Now, allow the total number of groups, J, to exceed two, and define d^* and d as

$$\begin{aligned} & d^* = z\delta + v, \\ & d = j \ \text{ if } \ C_{j-1} < d^* \leq C_j, \quad j=1,\ldots,J, \\ & \text{where } \ C_0 = -\infty, \ \text{ and } \ C_J = \infty, \end{aligned} \tag{12}$$

where now z does not contain an intercept. Such a model is applicable when, for example, the goal is to estimate women's labor force participation or choice of the employment type (wage-employment versus self-employment) by education level.

When specifying the joint distribution of u_j and v, it is convenient to assume normality for both errors, so that (12) becomes an ordered probit model. Similar to the two-group case, accounting for self-selection or, generally, sorting into groups is necessary when the correlation between u_j and v is different from zero.

Formally, let Assumption 2.1 hold for $j=1,2,\ldots,J$. Then, using the argument similar to the one in Section 2.1,

$$\begin{aligned} & y_j = 1[x\beta_j + \rho_j v + e_j > 0], \qquad j=1,\ldots,J, \\ & e_j|z,v \sim Normal(1,1-\rho_j^2). \end{aligned} \tag{13}$$

From (12) and (13), the conditional probabilities for each group are given by

$$\begin{aligned}
\mathrm{P}(y=1|d=1,z) &= \frac{\int_{-\infty}^{C_1-z\delta}\Phi\left(\frac{x\beta_1+\rho_1 v}{\sqrt{1-\rho_1^2}}\right)\phi(v)dv}{\Phi(C_1-z\delta)}, \qquad (14)\\
\mathrm{P}(y=1|d=j,z) &= \frac{\int_{C_{j-1}-z\delta}^{C_j-z\delta}\Phi\left(\frac{x\beta_j+\rho_j v}{\sqrt{1-\rho_j^2}}\right)\phi(v)dv}{\Phi(C_j-z\delta)-\Phi(C_{j-1}-z\delta)}, \quad j=2,\ldots,J-1,\\
\mathrm{P}(y=1|d=J,z) &= \frac{\int_{C_{J-1}-z\delta}^{\infty}\Phi\left(\frac{x\beta_J+\rho_J v}{\sqrt{1-\rho_J^2}}\right)\phi(v)dv}{1-\Phi(C_{J-1}-z\delta)}.
\end{aligned}$$

By differentiating the conditional probability with respect to x_k, the partial effects of continuous explanatory variables are obtained as

$$\begin{aligned}
PE_{j,k} &= \frac{\delta_k}{\Phi(\alpha_j)-\Phi(\alpha_{j-1})}\cdot\left[\phi(\alpha_{j-1})\Phi\left(\frac{x\beta_j+\rho_j\alpha_{j-1}}{\sqrt{1-\rho_j^2}}\right)-\phi(\alpha_j)\Phi\left(\frac{x\beta_j+\rho_j\alpha_j}{\sqrt{1-\rho_j^2}}\right)\right]\\
&+ \frac{1}{\Phi(\alpha_j)-\Phi(\alpha_{j-1})}\cdot\frac{\beta_{jk}}{\sqrt{1-\rho_j^2}}\cdot\int_{-\Phi(\alpha_{j-1})}^{\Phi(\alpha_j)}\phi\left(\frac{x\beta_j+\rho_j v}{\sqrt{1-\rho_j^2}}\right)\phi(v)dv\\
&+ \delta_k\cdot\frac{\phi(\alpha_j)-\phi(\alpha_{j-1})}{\Phi(\alpha_j)-\Phi(\alpha_{j-1})}\cdot\mathrm{P}(y=1|d=j,z), \qquad (15)
\end{aligned}$$

for $j=1,\ldots,J$, $C_0=-\infty$, $C_J=\infty$, $\alpha_j=C_j-z\delta$, and $\mathrm{P}(y=1|d=j,z)$ as defined in equation (14). Partial effects of discrete variables are obtained as in equation (11), using the conditional probabilities in (14).

2.3 Model for unordered multiple groups

In some cases we can have multiple groups that are not ordered. For example, one might want to study the determinants of job satisfaction among workers employed in different types of jobs: wage-employment in the private sector, wage-employment in the public sector, and self-employment. In such a case, the choice of $d=j$ can be described in the

context of a multinomial response model. To formalize ideas, define

$$d_j^* = z\delta_j + v_j, \quad j = 1, \ldots, J, \tag{16}$$

where the error now varies by group.

Following the standard formulation of a multinomial response model, the cross-section unit i will be in group j if it has the highest chance of belonging to that group. In the case of self-selection, choice j is the best option in the available set. We can write it as

$$d = j \text{ if } d_j^* = \max\{d_1^*, d_2^*, \ldots, d_J^*\}, \tag{17}$$

where the choice in equation (17) will be made if $z\delta_j + v_j > z\delta_l + v_l$ for all $l \neq j$. It is clearly seen that only differences between d_j^* are identified, so that a reference category needs to be assigned – a feature that is common for all multinomial response models.

To obtain the formulae for the conditional probability of y, we define $\tilde{v}_l = v_j - v_l$, $\tilde{\delta}_l = \delta_j - \delta_l$, for $l \neq j$, and make the following assumption:

ASSUMPTION 2.3.1 (i) $(u_j, v_1, \ldots, v_J)$ are independent of (x, z), $j = 1, \ldots, J$, (ii) $z = (x, z_1)$, where z_1 is not empty, (iii) $(u_j, \tilde{v}_1, \ldots, \tilde{v}_{j-1}, \tilde{v}_{j+1}, \ldots, \tilde{v}_J)$ have a multinomial normal distribution, $j = 1, \ldots, J$, (iv) $0 < \mathrm{P}(d = j) < 1$, $j = 1, \ldots, J$.

Under Assumption 2.3.1, for group $j = 1$, for example, we obtain

$$\begin{aligned} P(y = 1, d = 1) &= \int_{-x\beta_1}^{\infty} \int_{-z\tilde{\delta}_2}^{\infty} \cdots \int_{-z\tilde{\delta}_J}^{\infty} \phi(u_1, \tilde{v}_2, \ldots, \tilde{v}_J) du_1 d\tilde{v}_2, \ldots, d\tilde{v}_J, \\ P(d = 1) &= \int_{-z\tilde{\delta}_2}^{\infty} \cdots \int_{-z\tilde{\delta}_J}^{\infty} \phi(\tilde{v}_2, \ldots, \tilde{v}_J) d\tilde{v}_2, \ldots, d\tilde{v}_J, \end{aligned} \tag{18}$$

and the conditional probability is obtained as $P(y = 1|d = 1) = \frac{P(y=1,d=1)}{P(d=1)}$. Probabilities $P(y = 1|d = j)$, $j = 2, \ldots, J$, are obtained similarly.

Because equation (18) does not have a closed form solution, one would need to nu-

merically evaluate this J-dimensional integral. While simulated likelihood methods have been helpful in addressing computational difficulties, the estimation may still be infeasible if there are more than four groups. Therefore, we also consider a different approach.

The unordered multiple groups case can be considered in the context of selection models, where the choice is made between the best option and the second best alternative. This approach appears to be well-suited for the presented model because when estimating group-specific effects, one needs to account for the fact that option $d = j$ is chosen as the most optimal of all. We can define a binary indicator for group j as

$$\begin{aligned} w_j &= 1[z\delta_j + v_j > \bar{d}_j], \\ \bar{d}_j &= \max_{l \neq j}\{z\delta_l + v_l\}, \end{aligned} \tag{19}$$

which can be re-written as

$$w_j = 1[z\bar{\delta}_j + \bar{v}_j > 0], j = 1, \ldots, J, \tag{20}$$

where $\bar{\delta}_j$ is a difference between δ_j and the vector of parameters that correspond to $\bar{d}_j$, and $\bar{v}_j$ is a difference between v_j and the error corresponding to $\bar{d}_j$. Because in the unordered case the second best option is not known, $\bar{\delta}_j$ is determined as a weighted average of $\delta_j - \delta_l$, $l \neq j$, where weights depend on the probability that group l is the best alternative to j.

In this case, we will formulate the assumption for the errors in equation (20), rather than v_j in (19). Specifically, assume

ASSUMPTION 2.3.2 (i) $(u_j, \bar{v}_j)$ are independent of (x, z), $j = 1, \ldots, J$, (ii) $z = (x, z_1)$, where z_1 is not empty, (iii) $(u_j, \bar{v}_j)$ have a bivariate normal distribution with $\mathrm{Var}(u_j) = \mathrm{Var}(\bar{v}_j) = 1$ and $\mathrm{Corr}(u_j, \bar{v}_j) = \rho_j$, $j = 1, \ldots, J$, (iv) $0 < \mathrm{P}(d = j) < 1$, $j = 1, \ldots, J$.

Notice that in this model it is not possible to estimate δ_j. Fortunately, this does not affect our ability to consistently estimate parameters β, which is the main goal of the

estimation.

Under Assumption 2.3.2, conditional probabilities for each group are obtained as

$$\mathrm{P}(y=1|d=j,z)=\mathrm{P}(y=1|w_j=1,z)=\frac{\int_{-\infty}^{z\bar{\delta}_j}\Phi\left(\frac{x\beta_j+\rho_j\bar{v}}{\sqrt{1-\rho_j^2}}\right)\phi(\bar{v})d\bar{v}}{\Phi(z\bar{\delta}_j)},\quad j=1,\ldots,J. \tag{21}$$

Moreover, the partial effect of a continuous variable x_k is given by

$$\begin{aligned}PE_{j,k}&=\frac{\partial\mathrm{P}(y=1|d=j,z)}{\partial x_k}=\bar{\delta}_k\cdot\frac{\phi(z\bar{\delta})}{\Phi(z\bar{\delta})}\cdot\Phi\left(\frac{x\beta_j+\rho_j z\bar{\delta}}{\sqrt{1-\rho_j^2}}\right) \\ &+\frac{1}{\Phi(z\bar{\delta})}\cdot\frac{\beta_{jk}}{\sqrt{1-\rho_j^2}}\cdot\int_{-\infty}^{z\bar{\delta}}\phi\left(\frac{x\beta_j+\rho_j\bar{v}}{\sqrt{1-\rho_j^2}}\right)\phi(\bar{v})d\bar{v}-\bar{\delta}_k\cdot\frac{\phi(z\bar{\delta})}{\Phi(z\bar{\delta})}\cdot\mathrm{P}(y=1|w_j=1,z).\end{aligned} \tag{22}$$

Similar to the previous discussion, partial effects of discrete covariates are obtained as differences in conditional probabilities.

3 Estimation

To estimate the models presented in Section 2, one can use the maximum likelihood (MLE) estimator. In the case of two groups, the switching probit estimator is obtained by maximizing the following log likelihood function:

$$\begin{aligned}\ln L &= \sum_{i=1}^{N}\ln L_i, \\ L_i &= P_{i,11}^{y_{i1}(2-d_i)}\cdot P_{i,01}^{(1-y_{i1})(2-d_i)}\cdot P_{i,12}^{y_{i2}(d_i-1)}\cdot P_{i,02}^{(1-y_{i2})(d_i-1)}, \\ P_{i,11} &\equiv P(y_i=1,d_i=1|z_i)=\int_{-\infty}^{-z_i\delta}\Phi\left(\frac{x_i\beta_1+\rho_1 v}{\sqrt{1-\rho_1^2}}\right)\phi(v)dv, \\ P_{i,01} &\equiv P(y_i=0,d_i=1|z_i)=\int_{-\infty}^{-z_i\delta}\left[1-\Phi\left(\frac{x_i\beta_1+\rho_1 v}{\sqrt{1-\rho_1^2}}\right)\right]\phi(v)dv, \\ P_{i,12} &\equiv P(y_i=1,d_i=2|z_i)=\int_{-\infty}^{z_i\delta}\Phi\left(\frac{x_i\beta_2+\rho_2 v}{\sqrt{1-\rho_2^2}}\right)\phi(v)dv,\end{aligned} \tag{23}$$

$$P_{i,02} \equiv P(y_i = 0, d_i = 2|z_i) = \int_{-\infty}^{z_i\delta} \left[1 - \Phi\left(\frac{x_i\beta_2 + \rho_2 v}{\sqrt{1-\rho_2^2}}\right)\right] \phi(v)dv.$$

The resulting estimators of β_1, β_2, ρ_1, and ρ_2 are unbiased under the standard set of assumptions for the MLE estimator. Then, it would be possible to check whether the usual group-by-group estimation is valid by testing the hypothesis, $H_0 : \rho_1 = \rho_2 = 0$. Also, the equality of the coefficients in the two groups could be tested either for each explanatory variable separately, or for the entire vectors of parameters, β_1 and β_2.

When specifying the likelihood function for $J > 2$, ordered groups, it is convenient to use indicators s_j, $j = 1, \ldots, J$, that were defined at the beginning of Section 2. When groups are ordered, the likelihood function for observation i can be written as

$$L_i = P_{i,11}^{y_{i1}s_{i1}} \cdot P_{i,01}^{(1-y_{i1})s_{i1}} \cdot \ldots \cdot P_{i,1J}^{y_{iJ}s_{iJ}} \cdot P_{i,0J}^{(1-y_{iJ})s_{iJ}}, \tag{24}$$

$$P_{i,11} \equiv \mathrm{P}(y_i = 1, d_i = 1|z_i) = \int_{-\infty}^{C_1 - z_i\delta} \Phi\left(\frac{x\beta_1 + \rho_1 v}{\sqrt{1-\rho_1^2}}\right) \phi(v)dv,$$

$$P_{i,01} \equiv \mathrm{P}(y_i = 0, d_i = 1|z_i) = \int_{-\infty}^{C_1 - z_i\delta} \left[1 - \Phi\left(\frac{x\beta_1 + \rho_1 v}{\sqrt{1-\rho_1^2}}\right)\right] \phi(v)dv,$$

$$P_{i,1j} \equiv \mathrm{P}(y_i = 1, d_i = j|z_i) = \int_{C_{j-1} - z_i\delta}^{C_j - z_i\delta} \Phi\left(\frac{x\beta_j + \rho_j v}{\sqrt{1-\rho_j^2}}\right) \phi(v)dv, \ j = 2, \ldots, J-1,$$

$$P_{i,0j} \equiv \mathrm{P}(y_i = 0, d_i = j|z_i) = \int_{C_{j-1} - z_i\delta}^{C_j - z_i\delta} \left[1 - \Phi\left(\frac{x\beta_j + \rho_j v}{\sqrt{1-\rho_j^2}}\right)\right] \phi(v)dv, \ j = 2, \ldots, J-1,$$

$$P_{i,1J} \equiv \mathrm{P}(y_i = 1, d_i = J|z_i) = \int_{C_{J-1} - z_i\delta}^{\infty} \Phi\left(\frac{x\beta_J + \rho_J v}{\sqrt{1-\rho_J^2}}\right) \phi(v)dv,$$

$$P_{i,0J} \equiv \mathrm{P}(y_i = 0, d_i = J|z_i) = \int_{C_{J-1} - z_i\delta}^{\infty} \left[1 - \Phi\left(\frac{x\beta_J + \rho_J v}{\sqrt{1-\rho_J^2}}\right)\right] \phi(v)dv.$$

Similar to the ordered probit model, maximization of the log-likelihood function is performed with respect to β_1, ..., β_J, ρ_1, ..., ρ_J, C_1, ..., C_{J-1}. Because all parameters are estimated together, it is easy to test the hypotheses involving parameters from different groups (e.g. testing parameter equality across j). This can be done using the usual t and

Wald tests, as well as the likelihood ratio test.

It is important to note that when $J = 2$, or if there are $J > 2$ ordered groups, it is possible to consistently estimate parameters separately for each group by specifying $P(d_i = l)$ rather than $P(y_i, d_i = l)$ for groups $l \neq j$. This might make it easier to perform the optimization and should not result in an efficiency loss, because different groups are never observed together. Of course, an important disadvantage is that testing the equality of parameters in different groups becomes more complicated.

For the unordered multiple groups, one can make a joint normality assumption for $(u_j, \tilde{v}_1, \ldots, \tilde{v}_{j-1}, \tilde{v}_{j+1}, \ldots, \tilde{v}_J)$, as in Assumption 2.3.1, and estimate parameters by MLE. Here, we focus on a simpler estimator that relies on Assumption 2.3.2, where estimation is similar to that discussed in the sample selection literature. The vectors of parameters are estimated separately for each group, where the estimator accounts for the choice of (or self-selection into) the group, which may be nonrandom.

For each group j, we specify the likelihood function as

$$L_i = P_{i,11}^{y_{i1}w_{ij}} \cdot P_{i,01}^{(1-y_{i1})w_{ij}} \cdot P_{i,0}^{(1-w_{ij})}, \tag{25}$$

where

$$\begin{aligned} P_{i,11} &\equiv P(y_i = 1, w_{ij} = 1|z_i) = \int_{-\infty}^{z_i\tilde{\delta}_j} \Phi\left(\frac{x_i\beta_j + \rho_j v}{\sqrt{1-\rho_j^2}}\right)\phi(v)dv, \\ P_{i,01} &\equiv P(y_i = 0, w_{ij} = 1|z_i) = \int_{-\infty}^{z_i\tilde{\delta}_j} \left[1 - \Phi\left(\frac{x_i\beta_j + \rho_j v}{\sqrt{1-\rho_j^2}}\right)\right]\phi(v)dv, \\ P_{i,0} &\equiv P(w_{ij} = 0|z_i) = 1 - \Phi(z_i\tilde{\delta}_j). \end{aligned} \tag{26}$$

The limitation of this estimation approach is that hypothesis testing is complicated when parameters from different groups are involved. A relatively simple solution is to use bootstrap, where all β_j are estimated using the same bootstrap samples. Then, it becomes

relatively easy to obtain covariance matrices for the estimators of slope parameters in different groups.

Partial effects are usually estimated using one of the two methods. First, it is possible to estimate average partial effects by averaging over the distribution of all covariates other then the one whose effect is being estimated. Alternatively, one can obtain partial effects evaluated at particular values of other explanatory variables, such as the sample mean or median values. In the empirical application below, we estimate average partial effects.

4 Extensions to Panel Data

To be added.

5 Empirical Application

To illustrate the presented theoretical argument with an empirical example, we study the determinants of labor force participation among white, African American, and Hispanic women. In this case, groups are defined by the person's race or ethnicity. In the literature, it is usually assumed that race/ethnicity can be viewed as exogenous because it is not a choice variable. However, there may be cultural and behavioral differences across different ethnic groups, which may also be related to economic outcomes, such as the probability of employment. For example, women in a particular ethnic group may traditionally be more independent, where independence may also impact the likelihood of working. Using the methodology presented above, it is possible to test this hypothesis, as well as study the sensitivity of estimation results to accounting for a possibility of nonrandom sorting.

To perform estimation, we use data from the National Longitudinal Survey of Youth, 1979 (NLSY79). The initial sample is representative of all individuals who were 14 to 22 years old in 1979. To maximize the sample size we use data from the 1990 wave of the

survey, where the response rate was relatively high (about 91%), and all supplemental samples (poor white, black, and Hispanic) were still active. In 1990, all respondents were at least 25 years old, and the age of the oldest respondent was 33. Women in the military sample and those working in a family business were excluded. After dropping the observations with missing information on any of the variables used in the analysis, the final sample includes 4,417 women, 2,585 of whom are white, 1,139 African American, and 693 Hispanic.

The dependent variable is an indicator equal to one if the woman worked for at least some time during the period since the last interview. The list of explanatory variables includes age, education and marital status indicators, number of young children (ages 0 through 5), number of older children (ages 6 through 17), income of the spouse (in thousands of dollars), urban location indicator, and region indicators. To control for individual differences in cognitive ability we include the woman's score on the Armed Forces Qualification Test (AFQT), which was administered in 1979. The AFQT score was standardized to have a zero mean and unit variance in the sample.

As mentioned earlier, it is necessary to have an exclusion restriction to ensure the reliability of the estimator. Such a restriction can be obtained by assuming that the probability of working is determined by economic factors (such as skills and educational qualifications, availability of other sources of income, presence of children), but personality traits may be of minor or no importance. On the other hand, there may be personality differences by race and ethnicity, which may emerge due to cultural and social factors. In the context of the presented analysis, we assume that self-esteem varies by race/ethnicity, but does not affect the probability of employment. The self-esteem measure, developed by Rosenberg (Rosenberg, 1965), is aimed to assess the degree of approval or disapproval toward oneself. In the sample, the self-esteem measure is standardized to have a zero mean and unit variance. Later we check the validity of the exclusion restriction by estimating the unrestricted model and testing the statistical significance of self-esteem measure in

each equation.

Summary statistics are presented in Table 1. As seen in the Table, white women are slightly more likely to be working and be married, but tend to have fewer kids of ages 6-17. Among the three groups of women, African American females are lease likely to be married and tend to have more older children. Among married women, the income of the spouse tends to be the highest among white respondents and lowest among African American women. With regard to location, African American females are more likely to live in the South, while Hispanic women mostly reside in urban locales. Finally, white women tend to have the highest AFQT scores, while the self-esteem score tends to be the lowest among Hispanic women.

To obtain main results, the employment equation was first estimated separately for each racial/ethnic group by probit. Subsequently, the same equations were estimated using the methodology described in Section 2.3. Relying on Assumption 2.3.2, estimation was performed using the MLE estimator with the likelihood function as defined in equations (25) and (26).

Results are presented in Table 2. It is apparent that the correlation between the errors in the main and group choice equations is the strongest among white women. The correlation is approximately -0.55 and is highly statistically significant, suggesting that the unobservables that determine the probability of working among white women are negatively correlated with the unobserved determinants of the likelihood of being white. One possible example of such an unobserved factor could be the traditional views of the women's role as housekeepers and care providers. The traditional views are likely to negatively impact the probability of employment, and may be more common to white women. The error correlation for African American and Hispanic women is smaller in magnitude and not statistically significant. This may be due to the lesser importance of traditional views for the employment outcome among nonwhite women because their decision to work may be largely dictated by economic necessity. For example, because

Table 1: Summary Statistics.

	White	Black	Hispanic
	(1)	(2)	(3)
Proportion working	0.84	0.81	0.80
	(0.36)	(0.40)	(0.40)
Age	29.24	29.05	28.96
	(2.26)	(2.23)	(2.29)
12 years of schooling (proportion)	0.43	0.43	0.43
	(0.50)	(0.49)	(0.49)
13-15 years of schooling (proportion)	0.20	0.28	0.27
	(0.40)	(0.45)	(0.44)
16 or more years of schooling (proportion)	0.25	0.14	0.10
	(0.43)	(0.34)	(0.30)
Proportion married	0.61	0.26	0.52
	(0.49)	(0.44)	(0.50)
Number of children ages 0-5	0.63	0.59	0.71
	(0.79)	(0.79)	(0.81)
Number of children ages 6-17	0.62	0.99	0.86
	(0.92)	(1.09)	(1.06)
Income of Spouse (in $1,000)	40.43	30.07	34.96
	(27.78)	(28.35)	(24.16)
Urban location (proportion)	0.72	0.83	0.94
	(0.45)	(0.37)	(0.24)
Northeast region (proportion)	0.20	0.14	0.15
	(0.40)	(0.35)	(0.36)
Northcentral region (proportion)	0.30	0.19	0.09
	(0.46)	(0.39)	(0.28)
South region (proportion)	0.33	0.60	0.30
	(0.47)	(0.49)	(0.46)
West region (proportion)	0.17	0.07	0.46
	(0.37)	(0.25)	(0.50)
AFQT score (standardized)	0.42	-0.62	-0.38
	(0.97)	(0.69)	(0.81)
Self-esteem score (standardized)	0.04	0.06	-0.18
	(1.02)	(0.98)	(0.97)
Observations	2,585	1,139	693

Proportions may not add up to one due to rounding.

Table 2: Estimated Partial Effects on the Probability of Being Employed.

	White women		Black women		Hispanic women	
	Single probit (1)	Joint estimation (2)	Single probit (3)	Joint estimation (4)	Single probit (5)	Joint estimation (6)
Age	-0.0052^{*}	-0.0058	-0.0045	-0.0035	0.0066	0.0059
12 years of sch.	0.0623^{***}	0.0611^{***}	0.1532^{***}	0.1307^{**}	0.0869^{**}	0.0803^{**}
13-15 years of sch.	0.0932^{***}	0.0807^{***}	0.1777^{***}	0.1506	0.1050^{**}	0.0971^{**}
$\geq$ 16 years of sch.	0.1120^{***}	0.1043^{***}	0.1903^{***}	0.1605	0.1084	0.1007
Married	0.0297^{*}	0.0348	0.1470^{***}	0.1255^{**}	0.0146	0.0134
# young children	-0.1062^{***}	-0.1233^{***}	-0.0850^{***}	-0.0650^{***}	-0.1235^{***}	-0.1141^{***}
# older children	-0.0439^{***}	-0.0484^{***}	-0.0291^{***}	-0.0215^{***}	-0.0475^{***}	-0.0429^{***}
Spouse's income	-0.0013^{***}	-0.0014^{***}	-0.0011^{*}	-0.0008	0.0002	0.0002
Urban location	0.0144	0.0013	0.0710^{**}	0.0589	0.1117^{**}	0.0992
AFQT score	0.0271^{***}	0.0566	0.1122^{***}	0.0789^{***}	0.0838^{***}	0.0752^{**}
Corr$(u_j, \bar{v}_j)$		-0.5537^{***}		-0.3081		-0.1012
Observations	2,585	4,417	1,139	4,417	693	4,417

Statistical significance corresponds to the test of the underlying coefficient being equal to 0.
All equations also include region indicators.

African American and Hispanic women tend to have more children, the stay-at-home option may appear to be more economically viable for these women. On the other hand, non-economic factors may be less relevant.

Comparing the magnitude of the estimated partial effects produced by different estimation methods, it is seen that accounting for nonrandom group sorting alters the estimates somewhat. The most noticeable changes are observed for African American women. The estimated effects of most variables decline in magnitude and in same cases become insignificant. These include several education indicators, spousal income, and urban indicator. For white women, the largest changes are observed for the urban location, number of young children, and AFQT score. Accounting for nonrandom sorting has a very minor influence on the estimates for Hispanic women.

Going back to the underlying assumptions, they included the requirement that the

sorting equation include at least one variable that affects sorting, but not the main outcome. In the presented analysis, self-esteem was used as such a factor. Indeed, the self-esteem score was highly significant in the sorting equations for all three racial/ethnic groups. Moreover, when the unrestricted model was estimated, the self-esteem measure remained highly statistically significant in each group sorting equation, but was highly insignificant in the employment equations. Thus, the employed exclusion restriction appears to be valid.

6 Conclusion

This paper discusses the methodology for consistently estimating heterogeneous parameters in binary response models. In addition to a two-group case, we consider estimating parameters for multiple heterogeneous groups, which may be ordered or unordered. As an illustration, we estimate heterogeneous effects on women's employment outcomes using NLSY79 data. We find that although accounting for nonrandom group sorting does not appear to matter in some groups, in several cases it produces notably different results as compared to the simple group-by-group estimation.

References

Basu, Anirban, 2014, Estimating Person-Centered Treatment (PeT) Effects Using Instrumental Variables: An Application to Evaluating Prostate Cancer Treatments. *Journal of Applied Econometrics* 29, 671-691.

Bjorklund and Moffitt, 1987, The Estimation of Wage Gains and Welfare Gains in Self-Election Models. *Review of Economics and Statistics* 69(1), 42-49.

Carrasco, Raquel, 1999, Transitions to and from Self-Employment in Spain: An Empirical Analysis. *Oxford Bulletin of Economics and Statistics* 61(3), 315-41.

Goldfeld, S.M. and R.E. Quandt, 1973, *Nonlinear Methods in Econometrics*. Amsterdam: North Holland.

Heckman, James J., 1979, Sample Selection Bias as a Specification Error. *Econometrica* 47(1), 153-61.

Heckman, James J., Sergio Urzua and Edward Vytlacil, 2006, Understanding Instrumental Variables in Models with Essential Heterogeneity. *Review of Economics and Statistics* 88(3), 389-432.

Kyriazidou, E., 1997, Estimation of a panel data sample selection model. *Econometrica* 65, 1335-1364.

Lee, L.F., 1978, Unionism and Wage Rates: A Simultaneous Equation Model with Qualitative and Limited Dependent Variables. *International Economic Review* 19, 415-433.

Maddala G.S. and F. Nelson, 1975, Switching Regression Models with Exogenous and Endogenous Switching. *Proceedings of the American Statistical Association* (Business and Economics Section), 423-426.

Maddala, G.S., 1983, *Limited Dependent Variable and Qualitative Variables in Econometrics*. Cambridge, U.K.: Cambridge University Press.

Manski, C., D. Sandefur, S. McLanahan, and D. Powers, 1992, Alternative Estimates of Family Structure During Adolescence on High School Graduation, *Journal of the American Statistical Association* 87, 25-37.

Newey, W.K., 2009, Two-step series estimation of sample selection models. *Econometrics Journal* 12, S217-S229.

Semykina, A. and J.M. Wooldridge, 2017, Binary Response Panel Data Models with Sample Selection and Self Selection. *Journal of Applied Econometrics*, forthcoming.

Vella, Frank, 1988, Generating Conditional Expectations from Models with Selectivity Bias. *Economics Letters* 28, 97-103.

Wooldridge, J.M., 1995, Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions. *Journal of Econometrics* 68, 115–132.